

# THE DESIGN OF OPTIMUM FILTERS FOR QUANTIZING A CLASS OF NON BANDLIMITED SIGNALS<sup>†</sup>

Jamal Tuqan and P. P. Vaidyanathan

Department of Electrical Engineering 136-93  
California Institute of Technology, Pasadena, CA 91125, USA.  
tuqan@systems.caltech.edu, ppvnath@sys.caltech.edu

## ABSTRACT

We consider the efficient quantization of a class of *non bandlimited* signals, namely the class of discrete time signals that can be recovered from their decimated version. By definition, these signals are oversampled and it is reasonable to expect that we can reap the same benefits of well known efficient A/D conversion techniques. Indeed, by using appropriate multirate reconstruction schemes, we first show that we can obtain a great reduction in the quantization noise variance due to the oversampled nature of the signals. To further increase the effective quantizer resolution, noise shaping is introduced by optimizing linear time invariant (LTI) and linear periodically time varying (LPTV)<sub>M</sub> pre- and post filters around the quantizer. Closed form expressions for the optimum filters and the minimum mean squared error are derived for each case.

## 1. INTRODUCTION

If a discrete time signal  $x(n)$  is bandlimited to  $[-\pi/M, \pi/M]$ , it can be recovered from its decimated version  $x(Mn)$  by low pass filtering. Consider now the class of discrete time signals that can be modeled as the sum of the outputs of  $L < M$  interpolation filters, where  $M$  is the interpolation factor, as shown in Fig. 1. If a discrete time signal  $x(n)$  is obtained using the model in Fig. 1, it is natural to expect that it can be recovered from its decimated version,  $x(Mn)$ , even though it is in general *not bandlimited*. As a simple example, assume that  $x(n)$  satisfies the model of Fig. 1 with  $L = 1$ . If  $F_0(e^{j\omega})$  is a Nyquist( $M$ ) filter [1], then,  $x(Mn)$  is equal to  $y(n)$  and we have the relation  $x(n) = \sum_k x(kM)f_0(n - kM)$ . In other words,  $x(n)$

is completely defined by the samples  $x(Mn)$  even though the filter  $F_0(e^{j\omega})$  is not necessarily ideal. In a similar way, we can talk about a decimation system for the case where  $F_0(z)$  is not Nyquist( $M$ ) and for the case of  $L \neq 1$ . The details can be found in [2, 3].

The main issue in this paper is how to take advantage of the signal model (Fig. 1) in preparing a quantized or compressed version of  $x(n)$ . Our study is inspired by important concepts found in A/D conversion applications that exploit the bandlimited property of the signals to be quantized, such as, oversampling PCM techniques and noise

shaping. To elaborate more, consider the schematic shown in Fig. 3 where the box labeled  $Q$  is a simple uniform roundoff (PCM) quantizer. Assuming that the signal  $x(n)$  is bandlimited, we can low pass filter the quantized signal  $x(n) + e(n)$  where  $e(n)$  is the quantization noise. The ideal low pass filter on the right removes the noise in the stopband but does not change the signal component. The signal power remains unchanged whereas the noise power decreases proportionally to the oversampling ratio. By exploiting the oversampled nature of the signal  $x(n)$ , we can therefore trade off quantizer complexity for higher resolution. This technique is usually termed oversampled PCM conversion [4]. Consider now the system of Fig. 4 where  $P(e^{j\omega})$  is a linear time-invariant (LTI) filter. The input signal  $x(n)$  is still assumed to be bandlimited. In addition to the benefits described above, it can be shown that a clever choice of the filter  $P(e^{j\omega})$  in Fig. 4 produces a further decrease in the noise power. The filter  $P(e^{j\omega})$  affect the noise component  $e(n)$  but not the input signal  $x(n)$ . The system of Fig. 4 introduces *noise shaping* in the signal band to allow higher resolution quantization of bandlimited signals.

With these ideas in mind, observe now the output  $x(n)$  of Fig. 1. Even though  $x(n)$  is not bandlimited, it can be reconstructed from its downsampled version as explained previously. In this sense, *it can be considered as an oversampled signal*. The following questions then arise: Can we obtain the same advantages of the oversampling PCM conversion technique for a non bandlimited signal satisfying the model of Fig. 1? Furthermore, for a fixed set of filters  $F_k(e^{j\omega})$ ,  $k = 0, 1, \dots, L-1$ , what is the optimal filter  $P(e^{j\omega})$  that minimizes the noise power at the output? Since the signal  $x(n)$  is cyclo-widesense stationary of period  $M$  [(CWSS)<sub>M</sub>] [5], restricting ourselves to linear time invariant noise shaping filters is a loss of generality. So, by using a linear periodically time varying (LPTV)<sub>M</sub> scheme, can we decrease the error further? These are the questions we address and answer in this paper.

## 2. FILTER AND QUANTIZER ASSUMPTIONS

**Filter assumptions.** The FIR filters  $F_k(e^{j\omega})$  of Fig. 1 are assumed to be the synthesis filters corresponding to the first  $L$  channels of an  $M$ -channel *principal component* orthonormal filter bank. Although not necessary for the derivation of the results of this paper, the assumption is motivated by the following fact: Given a wide-sense stationary (WSS) signal with energy concentrated mostly in certain subbands, the problem of finding the best signal

<sup>†</sup> This work is supported in parts by the Office of Naval Research grant N00014-93-1-0231, Tektronix, Inc., and Rockwell International.

model in the mean squared sense reduces to that of finding the filter bank that produces the  $L$  most dominant subbands. If the filter bank is orthonormal, the modeling issue reduces to the design of a principal component filter bank ( $L \neq 1$ ) or the design of energy compaction filters ( $L = 1$ ). The design of FIR PCFB and FIR globally optimal energy compaction filters is discussed in [6, 7].

**Quantizer assumption.** The box labeled  $Q$  represents a scalar uniform (PCM) quantizer and is modeled as an additive zero mean white noise source  $q(n)$ . Because the model filters are not ideal, the input  $x(n)$  is a zero mean  $(CWSS)_M$  process. Since the input to the quantizer  $x(n)$  is a  $(CWSS)_M$  process, its variance  $\sigma_x^2(n)$  is a periodic function of  $n$  with period  $M$ . Define  $\sigma_x^2$  to be the average

variance of  $x(n)$ , i.e.,  $\sigma_x^2 = \frac{1}{M} \sum_{n=0}^{M-1} \sigma_x^2(n)$ . Then, choose

the fixed step size  $\Delta$  in the uniform quantizer such that the quantization noise variance  $\sigma_q^2$  is directly proportional to the average variance of the quantizer input  $x(n)$ , that is

$$\sigma_q^2 = c2^{-2b} \sigma_x^2 \quad (1)$$

where  $\sigma_q^2$  is the quantization noise variance,  $c$  is a constant that depends on the statistical distribution of  $x(n)$  and the overflow probability, and  $\sigma_x^2$  is the average variance of the quantizer input. The above relation is justified for a PCM quantizer using 3 (or more) bits per sample (see chapter 4 in [8]).

### 3. INCREASING THE QUANTIZER RESOLUTION BY MULTIRATE FILTERING

Consider the set up shown in Fig. 5. The tilde accent on a function  $F_k(e^{j\omega})$  is defined such that  $\tilde{F}_k(e^{j\omega})$  is the conjugate of  $F_k(e^{j\omega})$ . In the absence of quantization, the scheme is a perfect reconstruction system. In the presence of the quantizer, the output  $\hat{x}(n)$  in Fig. 5 is equal to the original sequence  $x(n)$  plus an error signal  $e(n)$  due to quantization. The following result shows that, by using the above scheme, a significant reduction in the average mean square

error  $\mathcal{E} \triangleq \frac{1}{M} \sum_{n=0}^{M-1} E\{e(n)\}^2$  can be obtained in comparison with the direct quantization of  $x(n)$  shown in Fig. 2.

**Theorem 1** Consider the scheme of Fig. 5 where the  $L$  filters  $F_k(e^{j\omega})$  are assumed to be any  $L$  channels of an  $M$ -channel critically sampled orthonormal filter bank. Under the above quantization noise assumption, the average mean square error  $\mathcal{E}$  is equal to  $\frac{L}{M} \sigma_q^2$ .

The proof can be found in [9]. The quantization noise variance  $\sigma_q^2$  obtained by directly quantizing  $x(n)$  as shown in Fig. 2 is now reduced by the oversampling factor  $M/L$ . The signal variance  $\sigma_x^2$  on the other hand did not change. We note the following:

1. The signal  $x(n)$ , modeled as in Fig. 1, is oversampled and therefore, contains redundant information in the form of an excess of samples. We are therefore effectively quantizing with a higher number of bits per sample. This trade off, between the quantization noise variance (effective quantizer

resolution) and the sampling rate is the underlying principle of oversampled A/D converters.

2. The parameter  $L$ , defined to be the number of channels, alternates between two extremes:  $L = 1$  and  $L = M$ . When  $L = 1$ , we get the best SNR improvement at the expense of a more narrow class of inputs  $x(n)$ . When  $L = M$ , no noise variance reduction is achieved since the class of signals is now unrestricted. We can also see this by noticing that the multirate interconnection in Fig. 5 becomes a perfect reconstruction filter bank that is signal independent. The parameter  $L$  therefore determines the tradeoff between the generality of the class of signals, modeled as in Fig. 1, and the reduction in quantization noise variance.

### 4. NOISE SHAPING BY LTI PRE- AND POST FILTERS

Following the philosophy of sigma-delta modulators, we now perform noise shaping to achieve a further reduction in the average mean square error. To accomplish this, we propose using LTI pre- and post filters around the PCM quantizer as shown in Fig. 6. The goal is to optimize these filters such that the average m.s.e. at the output is minimized. The noise shaping filters to be optimized are not constrained to be rational functions (i.e., of finite order). Non causal solutions, for example, are accepted.

Although our quantizer design assumptions are the same as before, the quantizer input is not anymore the  $(CWSS)_M$  process  $x(n)$ , but a filtered version of it, which we denote by  $z(n)$ . Following (1), the noise variance in this case is given by  $\sigma_q^2 = c2^{-2b} \sigma_z^2$  where  $\sigma_z^2$  is the average variance of the process  $z(n)$ . It is then possible to express  $\sigma_z^2$  in terms of the prefilter  $P(e^{j\omega})$  and the so called average power spectral density of the process  $x(n)$ , denoted by  $\hat{S}_{xx}(e^{j\omega})$ , as follows:

$$\sigma_z^2 = \frac{1}{M} \int_{-\pi}^{\pi} |P(e^{j\omega})|^2 \hat{S}_{xx}(e^{j\omega}) \frac{d\omega}{2\pi} \quad (2)$$

The average power spectral density is a familiar concept that arises when "stationarizing" a  $(CWSS)_M$  process [10] and satisfies the well known properties of the power spectrum of a WSS process. It is defined to be the discrete time Fourier transform of the time averaged autocorrela-

tion function  $\hat{R}_{xx}(k)$  given by  $\frac{1}{M} \sum_{n=0}^{M-1} E[x(n)x^*(n-k)]$ .

**Theorem 2** [9] Consider the scheme of Fig. 6 under the assumptions of section 2. The optimum filter  $P(e^{j\omega})$  that minimizes the average mean square reconstruction error has the following magnitude squared response:

$$|P_{opt}(e^{j\omega})|^2 = \frac{\sqrt{(\sum_{i=0}^{L-1} |F_i(e^{j\omega})|^2)}}{\sqrt{\hat{S}_{xx}(e^{j\omega})}} \quad (3)$$

For the case of  $L = 1$ , the expression becomes:

$$|P_{opt}(e^{j\omega})|^2 = \frac{|F(e^{j\omega})|}{\sqrt{\hat{S}_{xx}(e^{j\omega})}} \quad (4)$$

and can be regarded as a multirate extension of the half whitening filter [8]. The coding gain of Fig. 6, defined as

the ratio  $\mathcal{E}_{direct}/\mathcal{E}_{min}$ , where  $\mathcal{E}_{direct}$  is the mean squared error obtained in the direct quantization case (Fig. 2) can be derived under the assumption that the JWSS processes  $y_k(n)$ ,  $k = 0, 1, \dots, L-1$ , are uncorrelated:

$$\mathcal{G}_{opt} = \frac{M \int_{-\pi}^{\pi} \sum_{k=0}^{L-1} S_{y_k}(e^{j\omega}) |F_k(e^{j\omega})|^2 \frac{d\omega}{2\pi}}{\left( \int_{-\pi}^{\pi} \sqrt{\sum_{i=0}^{L-1} S_{y_i}(e^{j\omega}) |F_i(e^{j\omega})|^2} \sum_{n=0}^{L-1} |F_n(e^{j\omega})|^2 \frac{d\omega}{2\pi} \right)^2} \quad (5)$$

The coding gain expression for the  $L = 1$  case becomes

$$\mathcal{G}_{opt} = \frac{M \int_{-\pi}^{\pi} S_{yy}(e^{j\omega}) \frac{d\omega}{2\pi}}{\left( \int_{-\pi}^{\pi} \sqrt{S_{yy}(e^{j\omega})} \frac{d\omega}{2\pi} \right)^2} = M \mathcal{G}_{hw} \quad (6)$$

where  $\mathcal{G}_{hw}$  is the half whitening coding gain of the WSS process  $y(n)$  [8]. The factor  $M$  in (6) is due to the over-sampled nature of the signal  $x(n)$ . It is interesting to note that the noise shaping contribution to  $\mathcal{G}_{opt}$  in (6), which we denote by  $\mathcal{G}_{hw}$ , is exactly the coding gain we would obtain by half whitening the WSS process  $y(n)$  in the usual way. Due to space limitation, specific numerical examples are not included in this paper but can be found in [9, 11].

## 5. NOISE SHAPING BY $(LPTV)_M$ PRE- AND POST FILTERS

In this section, we consider using  $(LPTV)_M$  pre- and post filters instead of LTI ones surrounding a periodically time varying  $((PTV)_M)$  quantizer. Since the signal model  $x(n)$  is  $(CWSS)_M$ , restricting ourselves to linear time invariant noise shaping filters and quantizers is a loss of generality. Any optimum configuration for such processes should consist of  $(LPTV)_M$  filters surrounding a  $((PTV)_M)$  quantizer. Using some well known multirate results, it can be shown that this new quantization configuration is equivalent to an  $M$ -channel maximally decimated filter bank with  $M$  subband quantizers [1]. Because the general  $(LPTV)_M$  problem is difficult to track analytically, we will only study a special form of a perfect reconstruction filter bank. In specific, we assume that the analysis polyphase matrix  $\mathbf{E}(e^{j\omega})$  is diagonal with diagonal elements equal to  $V_k(e^{j\omega})$ . The synthesis polyphase matrix  $\mathbf{R}(e^{j\omega})$  is then also diagonal with diagonal elements equal to  $1/V_k(e^{j\omega})$  for each  $k$ . The quantization configuration is shown in Fig. 7. The scalar quantizers labeled  $Q$  are modeled as additive noise sources  $q_k(n)$  and individually satisfy relation the (1). We assume that the subband quantization noise sources  $q_k(n)$  are white and pairwise uncorrelated, i.e., the noise power spectral density matrix is given by

$$\mathbf{S}_{qq}(e^{j\omega}) = \begin{pmatrix} \sigma_{q_0}^2 & 0 & \dots & 0 \\ 0 & \sigma_{q_1}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & \sigma_{q_{M-1}}^2 \end{pmatrix} \quad (7)$$

The goal is then to jointly allocate the subband bits  $b_k$  under a fixed bit rate

$$b = \frac{1}{M} \sum_{k=0}^{M-1} b_k \quad (8)$$

and optimize  $V_k(e^{j\omega})$  in order to minimize the average m.s.e. at the output of Fig. 7.

**Theorem 3** Consider the scheme of Fig. 7 under the above assumptions. The optimum filter  $V_{opt,k}(e^{j\omega})$  (for each  $k$ ) that minimizes the average mean square reconstruction error at the output has the following magnitude squared response:

$$|V_{opt,k}(e^{j\omega})|^2 = \frac{\sqrt{\sum_{i=0}^{L-1} |\tilde{R}_{ik}(e^{j\omega})|^2}}{\sqrt{S_k(e^{j\omega})}} \quad (9)$$

where  $S_k(e^{j\omega}) = \sum_{i=0}^{L-1} S_{y_i}(e^{j\omega}) |\tilde{R}_{ik}(e^{j\omega})|^2$  is the power spectrum of  $k$ th channel and  $\tilde{R}_{ik}(e^{j\omega})$  is the  $k$ th polyphase component of the  $i$ th filter  $\tilde{F}_i(e^{j\omega})$ . Using the above optimum filters, the coding gain of Fig. 7 is then given by:

$$\mathcal{G}_{opt} = \frac{\sigma_y^2}{M \left( \prod_{k=0}^{M-1} \int_{-\pi}^{\pi} \sqrt{S_k(e^{j\omega})} \sqrt{\sum_{i=0}^{L-1} |\tilde{R}_{ik}(e^{j\omega})|^2} \frac{d\omega}{2\pi} \right)^{2/M}} \quad (10)$$

The case of  $L = 1$  yields again an interesting result.

**Theorem 4** Consider the scheme of Fig. 7 under the same above assumptions and with  $L = 1$ . The optimum filter  $V_{opt}(e^{j\omega})$  that minimizes the average mean square reconstruction error at the output is independent of  $k$  and has the following magnitude squared response:

$$|V_{opt}(e^{j\omega})|^2 = \frac{1}{\sqrt{S_{yy}(e^{j\omega})}} \quad (11)$$

where  $S_{yy}(e^{j\omega})$  is the power spectrum of the WSS process  $y(n)$ . With the above optimum filter expression, the coding gain is then given by:

$$\mathcal{G}_{opt} = \frac{\sigma_y^2}{M \left( \prod_{k=0}^{M-1} \int_{-\pi}^{\pi} \sqrt{S_{yy}(e^{j\omega})} |\tilde{R}_k(e^{j\omega})|^2 \frac{d\omega}{2\pi} \right)^{2/M}} \quad (12)$$

where  $\tilde{R}_k(e^{j\omega})$  is the  $k$ th polyphase component of  $\tilde{F}(e^{j\omega})$ .

**The LTI case is indeed a loss of generality.** Since the class of  $(LPTV)_M$  filters and  $(PTV)_M$  quantizers include the LTI case, it is clear that the performance of this more general class of filters and quantizers is at least as good as the LTI one. We have already shown that the optimum  $(LPTV)_M$  filter for the case of  $L = 1$  reduces to a LTI one. The question then becomes: Is the  $(PTV)_M$  quantizer providing any excess gain over the LTI case and if so, by how much? It turns out that, even in this restricted form of  $(LPTV)_M$  filters, the coding gain of the above scheme is always greater than the LTI one except when the magnitude squared response of the polyphase components  $R_k(e^{j\omega})$  of  $F(e^{j\omega})$  are equal for all  $k$ . Indeed, it can be shown [9] that, the denominator of (6) is always  $\geq$  than the denominator of (12) with equality if and only if all  $|\tilde{R}_k(e^{j\omega})|^2$  are equal. Since the numerator is the same in both cases, the claim is proved. For the more general case ( $L \neq 1$ ), we again expect the coding gain of the more general  $(LPTV)_M$  case of Fig. 7 to be higher than the analogous LTI one of Fig. 6. However, the complexity of the expressions (5) and (10) in this case prevents a formal mathematical proof.

## 6. REFERENCES

- [1] P. P. Vaidyanathan, *Multirate systems and filter banks*. Englewood Cliffs, NJ: Prentice Hall, 1993.
- [2] P. P. Vaidyanathan and S.-M. Phoong, "Reconstruction of sequences from non uniform samples," *ISCAS Proc.*, vol. 1, pp. 601-604, May 1995.
- [3] P. P. Vaidyanathan and S.-M. Phoong, "Discrete time signals which can be recovered from samples," *ICASSP Proc.*, vol. 2, pp. 1448-1451, May 1995.
- [4] P. Aziz, H. Sorensen, and J. V. der Spiegel, "An overview of sigma-delta converters," *IEEE signal processing magazine*, vol. 13, pp. 61-84, January 1996.
- [5] V. S. Sathe and P. P. Vaidyanathan, "Effects of multirate systems on the statistical properties of random signals," *IEEE Transactions on Signal Processing*, vol. 41, pp. 131-146, January 1993.
- [6] J. Tuqan and P. P. Vaidyanathan, "Globally optimal two-channel FIR orthonormal filter banks adapted to the input signal statistics," *ICASSP Proc.*, Seattle, May 1998.
- [7] J. Tuqan and P. P. Vaidyanathan, "The role of the discrete-time Kalman-Yakubovitch-Popov (KYP) lemma in designing statistically optimum FIR orthonormal filter banks," *ISCAS Proc.*, June 1998.
- [8] N. S. Jayant and P. Noll, *Digital coding of waveforms*. Englewood Cliffs, NJ: Prentice Hall, 1984.
- [9] J. Tuqan and P. P. Vaidyanathan, "Oversampling PCM techniques and optimum noise shapers for quantizing a class of nonbandlimited signals," *Submitted to the IEEE Transactions on Signal Processing*, December 1996.
- [10] W. A. Gardner, "Stationarizable random processes," *IEEE Transactions on Information Theory*, vol. 24, pp. 8-22, January 1978.
- [11] J. Tuqan and P. P. Vaidyanathan, "Optimum quantization of a class of non bandlimited signals," *Asilomar Conference on Signals, Systems, and Computers*, November 1996.

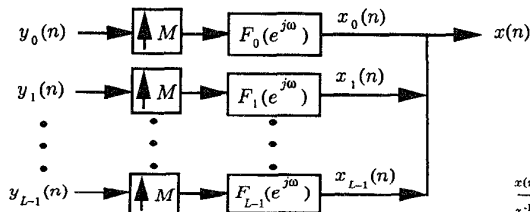


Figure 1: The multiband model

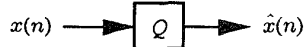


Figure 2: Direct quantization of  $x(n)$

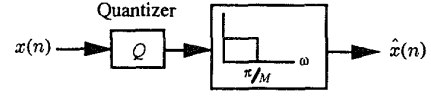


Figure 3: Schematic of the oversampling PCM technique

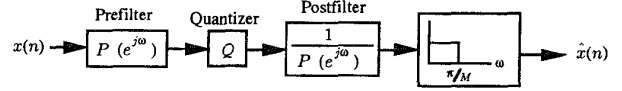


Figure 4: The quantization scheme of Fig. 3 with noise shapers

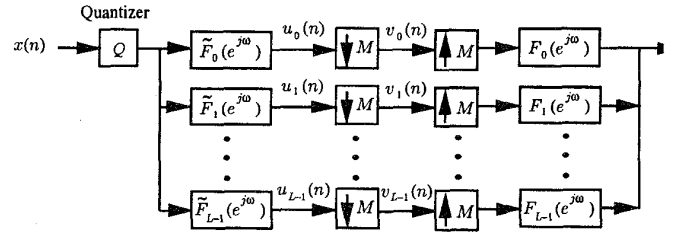


Figure 5: Multirate quantization scheme for the multiband model

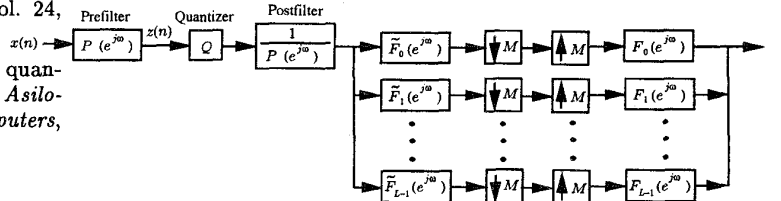


Figure 6: Noise shaping by LTI pre- and post filters for the multiband model

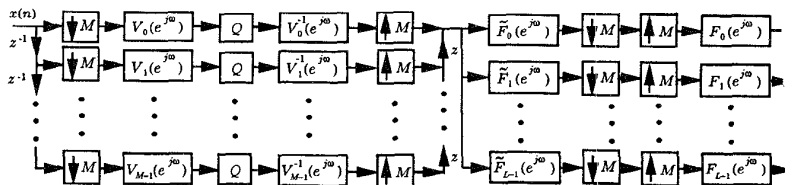


Figure 7: Noise shaping using  $(LPTV)_M$  pre- and post filters for the multiband model